Accepted version. Please cite from published version:

Taylor, C. 2018. Similarity. In Taylor, C. & A, Marchi (eds.). *Corpus Approaches to Discourse: A critical review*. Routledge.

Chapter 2: Similarity[1]

Charlotte Taylor, University of Sussex

2.1 Introduction

This aim of this paper is to raise the methodological importance of searching for similarity and stasis, as well as difference and change, in corpus-assisted discourse analysis. I aim to outline some of the possible methods for searching for similarity in corpus studies of discourse, and, more specifically, I look at methods which are accessible to language researchers who may not have a strong background in programing and/or statistics.

In the first section, a 'toolkit' is presented of the resources available to researchers looking for similarity. In the second section, a case-study is reported which employs some of the tools discussed in the previous section to investigate the similarity and stasis in the representation of *refugees* in British parliamentary and media discourse over the last 200 years.

In many ways, of course, corpus linguistics is founded on similarity, because it involves the search for 'usuality' and repeated patterns of behaviour. However, we are generally most used to focussing explicitly on similarity and comparability as key concepts at the stage where we are selecting or creating comparable corpora or reference/comparator corpora. For instance, Kilgarriff (2001) shows the importance of identifying similarity between and within (the homogeneity) corpora in order to assess the extent to which grammars or other tools may

[1] Parts of the overview presented in Sections 2.1-2.3 were first published in Taylor (2013).

usefully be extended from one to another. Likewise, Gries & Hilpert (2008) address similarity between corpus segments as a way of creating meaningful divisions in diachronic corpora (see also Chapter 10 of this volume).

In this chapter, I argue that the analysis for similarity can also be profitably expanded beyond the corpus selection/creation stage into the discourse analysis. Thus, this will be the kind of similarity which interests us here. Furthermore, I argue that this aspect is somewhat neglected within corpus/discourse studies (and indeed more generally but that is beyond the scope of this chapter). To take a brief snapshot of some recent work in corpus and discourse studies, I identified all articles published in *Corpora* and *International Journal of Corpus Linguistics* since 2015 which referred to discourse or stylistics in the title, keywords or abstract (28 out of a total of 55 published papers) and concordanced them for references to *same*/*similar** and *differen**. There were 784 occurrences of *differen** across the 28 texts, compared to 499 for *same*/*similar** across all 28 texts. This is a rather rough measure as it does not tell us the context of use, but it certainly points towards a concentration of attention on what is different. For a more accurate estimate, we could look at the research aims and what is presented as the key findings. In this case, the pattern was stronger with none exclusively focussing on similarity and eight only addressing difference. Although not represented in the journals examined here, one exception to this general pattern of backgrounding of similarity at the level of analysis is the area of authorship studies which systematically addresses similarity across texts for the purposes of attribution of texts to a particular author (mainly within forensic linguistics) and description of the style of a particular author (mainly within stylistics).

2.2 Why is similarity important for corpus & discourse work?

There are research questions we might want to pose which are entirely driven by an interest in similarity. For instance, in response to claims about the shifting of the centre ground to the

right in British politics, we might want to investigate whether political manifestoes show an increasing amount of similarity over time, converging towards the right-wing discourses. In addition, there are a set of reasons why looking at similarity can help us in more open questions posed as part of a project. For instance, if we were asking 'how is immigration represented in the British press?', rather than only looking at how it differs according to political orientation or newspaper type (broadsheet or tabloid) we might want to try and get a sense of any shared discourse patterns that characterise the UK press. Even with a difference-oriented starting point, such as 'how do student apology emails to female and male lecturers about differ?', as I will argue below, there is some benefit from deliberately addressing the opposite question 'and how are they the same?'.

I would like to suggest that there are several principal, inter-related reasons why we might also want to focus on similarity in corpus approaches to discourse studies. The first is simply that, by focussing on difference, we effectively create a 'blind spot'; this means that, rather than aiming for a 360-degree perspective of our data, we are actually starting out with the goal of achieving only a 180-degree visualisation. Therefore, the search for similarity can add a new range of starting points into our data and allow us to begin with a more ambitious aim regarding the 'completeness' of the analysis. In contrast, the potential neglect of patterns of similarities in the data leads to another significant threat to the balance of the analysis, which is that by setting out to look at difference, the analyst is likely to find and report on difference. No matter how arbitrary the construction of two corpora, if you carefully searched for differences between them, it is highly likely that you would be able to find some. Any such difference-oriented findings are potentially highly misleading as it may be that in quantitative terms the similarities between two corpora or topics considerably outweigh the differences. As Baker notes:

> [N]ot publishing or sharing such findings can result in what has been called 'bottom drawer syndrome'. For example, imagine that ten sets of researchers, working independently from each other, all build a corpus of Singapore English and compare it to a similar British corpus, looking at the same linguistic feature. In nine cases the researchers find that there are no significant differences, decide that the study is therefore uninteresting and assign the research to the bottom drawer of their filing cabinet rather than publishing it. However, the tenth researcher does find a difference and publishes the research, resulting in an inaccurate picture of what the general trend is when such a comparison is undertaken. (Baker, 2010: 83)

This, then, leads on to the second major motivation for highlighting the role of similarity, which is that the deliberate and systematic 'looking in both directions' may offer some kind of counter-balance to the issues of cognitive bias or corroboration drive. As Scott and Tribble (2006) have argued, we are pattern perceivers; indeed, 'it seems to be a characteristic of the homo sapiens [sic] mind that it is often unable to see things "as they are" but imposes on them a tendency, a trend, a pattern' (Scott and Tribble, 2006: 6). As they say, this insight and imagination has positive implications when manifested in our ability to identify patterns and is essential for discourse work, but there is also the risk of perceiving tendencies where there are none, or where they are the result of ignoring much of the data, as mentioned under the first point. The reporting of both difference and similarity could allow us, therefore, to check on unintentional bias and provide some evidence to counter any suspicions of intentional bias. Furthermore, there is the issue of the 'corroboration drive' (Marchi and Taylor, 2009), that is to say, the systematic search for elements that validate previous findings which is a variant of the more general confirmation bias. As researchers, we naturally tend towards building on our work; we look for corroboration that what we have found is valid and less frequently do we think to look for falsification or contrasting findings. Thus, a 'push' towards

also looking for what we are not expecting – similarity– may serve as a valuable check on that (natural) researcher instinct to focus on change and corroboration.

This is not an argument that difference and similarity will always be of equal importance, or that paying equal amounts of attention to each will automatically confer balance. But, at least, checking for the both will offer a more rounded view of the data. So, the search for similarity may help us to achieve a more complete picture of our data and, cumulatively, of our field of study; it helps counterbalance the issues of cognitive bias, and the reporting of similarity data provides robustness to the analysis. This search for similarity cannot, of course, remove the researcher from the research process, nor can it guarantee objectivity any more than any other form of triangulation; our corpus/discourse research will, therefore, naturally continue to be 'researcher-driven' (Taylor, 2010). As Stubbs reminds us, with reference to corpus stylistics, a purely automatic analysis is not possible because 'the linguist selects which features to study, the corpus linguist is restricted to features which the software can find, and these features still require a literary interpretation' (Stubbs 2005: 6). However, along with the key components of *transparency* and *replicability*, the search for similarity can help us to achieve a more methodologically sound analysis, not least because it pushes the software to find new features.

2.3 Similarity and the tools of corpus-assisted discourse research

The combination of corpus linguistics and (critical) discourse analysis provides the researcher with two potential macro starting points. In the first, the analyst may start from the corpus, making use of corpus software to access the data and identify for further exploration of any areas of interest, as may occur in an analysis which is driven by a keyword comparison. In the second, the analyst may start with a discourse-analytical frame and then use the corpus to collect data which is interpreted and categorised through that frame. In reality, of course, we are most likely to move or cycle between these different positions and perspectives at various

points in our research. However, for the purposes of this paper, I am primarily focussing on corpus analysis as the main entry point into the data.

The classic corpus linguistic entry points would include the analysis of concordances, collocates and keywords/key clusters and key semantic domains. In particular, I would argue that it is the popularity of the concept of keyness, and the provision of user-friendly software that can calculate keyness (e.g. AntConc, CQPWeb, Sketch Engine, WordSmith Tools), which has facilitated the analysis of difference. There are also tools which facilitate the comparison of collocates, such as Sketch Engines' Sketch Difference (discussed below) and the 'compare' function in the BYU interface to the British National Corpus (BNC; Davies, 2004) and other corpora. Both of these also allow for comparison across different corpora; for instance, it is noted on the BYU BNC page that, 'you can compare between registers – for example, verbs that are more common in legal or medical texts' (Davies, 2004), and this illustrates the (natural) emphasis on using the tools to search for difference.

Since these tools are accessible and very user-friendly, they constitute a prime example of how the tools which are available shape and form the type of research which may be carried out.[2] This is particularly the case for new researchers to an area, where they tend to start by learning the tools and then investigating questions which the tools facilitate. As McEnery and Hardie (2011: 42) note, 'if the toolset does not expand, then neither will the range of research questions that may be reasonably addressed using a corpus'. Thus it is hoped that the overt discussion of similarity in this chapter, like the topics in other chapters in this volume, may increase awareness both of the presence of these aspects and ways of looking into them from a corpus and discourse perspective. And, who knows, even to developments in the software to aid future investigations.

[2] For more on the relationship between tools and objects of observation see Partington (2009)

2.3.1 The habit of looking both ways

Although the focus here will be on the tools that can aid accounting for similarity in corpus and discourse work, it should be noted that a fundamental aspect of looking for similarity is by implementing the procedure of 'looking both ways' as a standard practice and methodological principle. For instance Seale et al. (2007) carry out a keyword comparison of broadsheet and tabloid corpora containing articles referring to sleep, and note that:

> A disadvantage lies in the fact that the method identifies differences between texts rather than similarities or overlaps which could be relevant. It was therefore important also to read and become familiar with the content of the articles and to use this knowledge to influence our interpretations. (Seale et al., 2007: 422)

Indeed, even if we think of *keyness* which is most often used to investigate difference, comparing two or more corpora against a reference corpus (rather than each other) would allow the researcher to identify similarities too.

In the following sub-sections, I start by considering what similarity-oriented analytical tools are embedded within existing software and then focus mainly on notions or procedures that have been developed within corpus linguistics for this purpose.

2.3.2 Collocate comparison

One of the tools mentioned in the previous section, Sketch Difference, allows us to analyse similarity as well as difference in collocational patterns because it includes shared collocates. For instance, Bednarek & Caple (2017) employ it to identify overlap in the use of *cyclist* and *cyclists* in the press and find that negativity is a shared news value associated with both forms. Sketch Difference can be used either to compare to words in the same corpus (as in the cyclist example), or the same word in two different (sub)corpora. This latter form of

comparison is illustrated in Figure 2.1, which displays part of the output for a comparison of *immigrant* in the 2005 and 2010 subcorpora of the SiBol British press corpus.[3]

[FIGURE 2.1 NEAR HERE]

The words which are coloured in grey (in colour originally) at the top of the image are those which are stronger collocates for the 2010 corpus, and those at the bottom in grey were strongest for the 2005 corpus. Those which are unshaded in the middle, are the shared collocates, thus they are the target for any study of similarity.

However, although there is this option of examining similarity at the same time, most research using Sketch Difference has so far focussed on the differences between items. As Pearce (2008: 21) notes in his study of the collocational behaviour of MAN and WOMAN, there is a risk that 'inevitably, with a tool (Sketch Difference) that is designed, as its name suggests, to reveal contrasts, the analyst is in danger of exaggerating the differences and overlooking similarities'.

Another Sketch Engine tool which can facilitate the analysis of similarity in terms of collocational patterns is the Thesaurus. The Sketch Engine Thesaurus is a distributional thesaurus which works by identifying the collocates of a word, and then in the second stage, identifies which other words share similar collocates term (see Rychlý and Kilgarriff 2007, for more detail on the algorithm used). So, for instance, if you were to look up *hot* in a general corpus, one of the highest ranking items the thesaurus is likely to produce is its antonym *cold*, because they occur in similar kinds of contexts (e.g. premodifying *water*, *weather* and so on). Figure 2.2 shows the findings for the Sketch Thesaurus for *immigrant* in EnTenTen13 (a 22-billion word web-based corpus collected in 2013). The size of the word in the visualisation corresponds to its ranking: the more similar the collocational patterns, the

[3] More information is available here: http://www.lilec.it/clb/?page_id=8

larger it is shown. Here we might note that it the apparent relational antonym of *emigrant* is not prominent while a semantic set relating to crime (*criminal, prisoner, offender, terrorist*) suggests similar lexical company is used.

[FIGURE 2.2 NEAR HERE]

2.3.3 Consistent collocates

Consistent collocates (or c-collocates) may be defined as 'words that stably collocate with the node in multiple datasets and are to be viewed as indicating core elements of meaning, semantic associations and semantic prosodies' (Germond, McEnery & Marchi 2016). The identification of consistent collocates seems to draw on Scott's (e.g., 1997) notion of consistency and the use of key-keywords (both discussed below), and was developed during work on the ESRC funded project 'Discourses of refugees and asylum seekers in the UK press, 1996–2006' project at Lancaster University, which was led by Paul Baker. The research team introduced this concept of consistent collocates (c-collocates) to describe the lexical items which collocated with *refugees* / *asylum seekers* / *immigrants* / *migrants* (RASIM) in at least seven out of the ten annual subcorpora which they had collected (described in Gabrielatos & Baker, 2008). The consistent collocates were calculated in order to exclude *seasonal collocates* – that is, words which may have been triggered by particular events, rather than being representative of newspaper discourse across the extended time period. Once again, to date, there seems to have been relatively uptake of this notion, although it has been more popular in diachronic studies (e.g. McEnery & Baker 2017).

2.3.4 Consistency analysis

WordSmith Tools allows for the creation of consistency lists when producing word lists, which will identify words which are shared across a number of texts. These consistency lists

are useful when working with several corpora, or corpora containing multiple files. According to Scott and Tribble (2006), the main uses are:

> First, in a general corpus like the BNC, to distinguish between wordtypes in terms of how consistently they get used in a mass of texts in the language. Second, if the scope of the research is the genre, to be able to locate lexical items which characterise certain genres or sub-genres. Third, to be able to study text variants (e.g., alternative translations or editions). (Scott and Tribble, 2006: 39)

Like keyword analysis, it requires the researcher to be working on a set of (sub)corpora. Scott (2001) employs the function in illustrating how a teacher of English for Specific Purposes (ESP) might identify core lexis by looking for items which occur across a number of relevant sub-corpora. A brief review of recent work suggests that this function is significantly underused compared to more popular tools such as keywords and collocates.

2.3.5 Key keywords

Key keywords are introduced in Scott (1997) and defined in the WordSmith Tools guide as follows: 'A "key key-word" is one which is "key" in more than one of a number of related texts. The more texts it is "key" in, the more "key key" it is' (Scott, 2016). Thus, while keywords identify what is different about one corpus compared with another, the analysis of key keywords allows the analyst to go on to identify how those differences and characterising features may be shared by other corpora – that is, to focus on similarity. Rather like the procedure for consistency analysis, key keywords are particularly useful when looking for repeated patterns across large numbers of sub-corpora. In Scott's (1997) model, the calculation of key keywords subsequently allows for the identification of associates, that is 'words found to be key in the same texts as a given key key word' (1997: 238), which form an alternative means of calculating collocation in the wider sense.

Although used more frequently than some other tools discussed here, as Bachmann (2011) notes there is still a scarcity of studies employing this procedure (Bachmann, 2011: 83). Bachmann draws on McEnery's (2009) use of keywords to identify transient and permanent key keywords in moral panic discourse (McEnery, 2009: 169) and applies them to his analysis of parliamentary debates in order to identify 'a list of concepts that are representative of the debates as a whole' (Bachmann, 2011: 87). In other words, it becomes a method for identifying concepts which are similar across the sub-corpora and is a means of avoiding isolated spikes of data. A similar notion is used in Fitzsimmons-Doolan (2009) in a study which functions as a model for this chapter as it both sets out to look for similarity and falsifies its own hypothesis – the hypothesis being that there would be similarities in the lexical patterns of newspaper discourse about language policies, and newspaper discourse about immigration. Fitzsimmons-Doolan uses WordSmith Tools for the calculation of keywords but adopts a manual analysis of what she calls the keyest keywords by counting: how many of the top 10 keywords from each corpus were (a) in the top 20 and (b) in the top 500 keywords lists for each of the other corpora. These measures show the distribution of the "keyest" keywords from one corpus within each of the other corpora. (Fitzsimmons-Doolan, 2009: 392) After the key keyword analysis, she found little similarity in terms of what characterised the sub-corpora of articles on language policies and articles on immigration.

2.3.6 Lockwords

Baker (2011) further addresses the issue of search for similarity by introducing the concept of *lockwords* which was designed to complement the existing notion of keywords by focussing on similarity in frequencies of lexical items across corpora. This notion, and the procedure used for determining them, was conceived as a result of his observations of stasis in the BLOB, LOB, FLOB and BE06 corpora. He notes that certain words 'were so consistent in their frequencies that they appeared to be the opposite of Scott's (2000) concept of keywords

–words which are highly frequent in one corpus when compared against another' (Baker, 2011: 73). Secondly, the notion of lockwords was conceived by Baker as a means of taking a more corpus-driven approach to diachronic language study, so that, rather than starting with a specific item or set of items to investigate, the researcher may start with lists of items that have or have not changed over the time period under study. One of the qualities of lockwords is that they may be used in conjunction with keywords as part of that principle of 'looking both ways', thus increasing the researcher's general awareness of patterns of both similarity and difference in two or more sets of corpora.

As the counterpart to keywords, lockwords may be relevant in most places that keywords are used, and yet they have seen surprisingly little uptake since 2011, both in terms of application and integration into existing software packages. At the time of writing, the lockword calculation is only available with CQPWeb (Hardie 2012) which calculates them using the log ratio method which is also applied to keywords and collocation.[4] Researchers using other software packages can calculate them manually, as detailed in Baker's (2011) original paper. In order to identify change in the corpora, Baker used the *WordSmith Tools* detailed consistency list (discussed above) to create lists of items for analysis and then calculated the coefficient of variance which is the ratio of the sample standard deviation to the sample mean. (Baker (2011: 72) notes that this 'is easily calculated by dividing the standard deviation by the mean and then multiplying by 100'.) This measure does not specify whether the change is an increase or decline, so in the final stage, the results need to be sorted manually.

2.3.7 Identifying the typical: ProtAnt

We might also consider *prototypicality* to be a measure of similarity. That is to say, the most prototypical text in a corpus is in some way the one that is most like the others. The

[4] See http://cass.lancs.ac.uk/?tag=cqpweb

identification of the most typical texts in a corpus is an important one for (critical) discourse studies. As researchers who combine both corpus linguistics and discourse analysis, we often need ways of identifying key texts for in-depth analysis. Although we often 'shunt' between the text and corpus level, entry point at text level can be, in itself, a form of methodological triangulation. The problem, as Anthony & Baker (2015) discuss in their presentation of a new software tool, ProtAnt (Anthony & Baker 2017) is how to identify texts for a (critical) discourse analysis starting point in a balanced and replicable way. They operationalise the concept of protypicality through ProtAnt which 'analyses the texts, generates a ranked list of keywords based on statistical significance and effect size, and then orders the texts by the number of keywords in them' (Anthony & Baker 2015: 274). Thus, the texts which are highest ranked are those with the highest number of keywords in them (compared to a reference corpus).

The concept of prototypicality, may also be used, like keyness itself, to get a sense of 'aboutness' (Scott & Tribble 2006) regarding the texts in a corpus. For instance, Bednarek & Caple's (2017) investigation of news values around cycling uses ProtAnt to identify the 'typical' values in each sub-corpus of newspapers from different countries. By focussing on the newspaper articles which were identified as most prototypical, they were able to look at which news values characterised the reporting across their sub-corpora. And, once again, they found that negativity was the key news value associated with cycling in the press. Thus we can see how tools not necessarily designed for investigating similarity may be 're-purposed' to fit this aim.

From this brief overview, we can see that, although they are not as prominent as the tools for searching for difference, we have access to a range of techniques for turning our focus

towards similarity. In the following case study, I will explore what results and picture may emerge when we focus on the search for similarity.

2.4 Case-study: Representation of refugees in political and media discourse

In this case study I set out to see what happens when we (re)focus our attention entirely towards similarity. Thus, this study is not intended to be representative of an entire research process, but rather a stage within a corpus/discourse analysis. The example on which I will concentrate is how the term *refugees* is used in the *Times* newspaper and British parliamentary debates over the last two hundred years.[5] In many diachronic corpus/discourse studies, the tendency is naturally to look at what changed over the time period and thus the risk is that we lose sight of what has remained constant over time. The example discussed here forms part of a wider project which sets out to address this imbalance and identify continuity in discourses of migration (understood as including both immigration and emigration) over the same time period. These two sources were chosen as they provide two potential avenues for similarity: 1) are there consistent patterns of representation in each corpus over time? 2) how similar are the representations in the law-making and reporting corpora, and does this relationship change over time?

2.4.1 The data

The two corpora used for this case-study are:

Times Online. This corpus was created at University of Lancaster, using the OCR (optical character recognition) files made available by the British Library. The corpus covers the period 1785-2011 (although, at the time of writing, the subcorpora for the decades 1910-1939 are not available). The current size is c. 8.5 billion words and it was analysed through

[5] The plural term was preferred because I was interested in collectivised representations.

CQPWeb. The scanned articles are also available to view as images through the Times Digital Archive to which many UK libraries subscribe.

Hansard Corpus. This resource was created by the SAMUELS consortium and hosted on the Brigham Young University corpus interface. [6] The corpus contains approximately 7.6 million parliamentary speeches from the period 1803-2005 and covers both the House of Commons and the House of Lords (overall size c.1.6 billion words). Access is available through the BYU corpus interface.

These two corpora represent an incredible resource for diachronic corpus-assisted discourse studies. However, there are also some challenges involved in using such huge resources hosted on (different) external websites. Namely, as they are discourse-complete corpora, and search-term specific, subcorpora made up of meaningful text units cannot be extracted in equivalent ways, which means that keywords and lockwords cannot be employed here. Thus, the main measure of similarity used in this study will be c-collocates. Furthermore, as the start and end dates do not match up precisely, for the purposes of comparison, only whole decades are included, thus the analysis covers 1810-2000.

2.4.2 Refugees

Of the many terms available for describing people who move across national boundaries (see, for instance, Gabrielatos & Baker 2008) *refugee* is perhaps one of the more sympathetic terms available, at least in the UK context. Indeed, if we consider recent debate about naming choices, much of it has centred around the apparent avoidance of the term *refugee* where it would be applicable, at least partly triggered by change in *Al Jazeera* editorial policy, made news in their article 'Why Al Jazeera will not say Mediterranean 'migrants'' (Malone 2015).

[6] For more on the SAMUELS consortium see http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/
For access to the corpus via the BYU interface see http://www.hansard-corpus.org/. Basic level access is free at the time of writing.

The fact that it is a more sympathetic term should also alert us to the fact that when we look at representation of *refugees*, we are not necessarily looking at the representation of the group of people who have the status of refugee according to the UN 1951 Refugee Convention. Indeed, in earlier work (Taylor 2014), I found that the contemporary British press tended to use *refugee* to refer to people who were forced to move elsewhere in the world. Those forced to move to the UK were either not discussed frequently or were described using another naming choice with a different set of connotations (e.g. *immigrant*).

Another feature that makes *refugees* an interesting lexical item for analysis in a study of similarity, is that it is a term that shows a high degree of similarity in frequency trends across the two corpora used here (discussed in the following section). Figure 2.3 shows a sample of possible naming choices that occur across the whole time period (thus excluding more recent names like *asylum seekers* or now archaic names like *aliens*). The naming choices tracked are: *immigrants*, *emigrants*, *foreigners* and *refugees*.

[FIGURE 2.3 NEAR HERE]

As can be seen, *emigrants* occurs much more frequently in the *Times* than in parliamentary discourse in the earlier stages, but both sources show a steep decline in more recent times (which does not reflect a simple absence of movement out of the UK). *Foreigners* shows a similar mismatch at the level of interest, which we could just attribute to size differences in the corpora, but this is not borne out by the closer frequencies in Hansard and *Times* for *immigrants* and *refugees* (both in the bottom half of the figure). For both *foreigners* and *immigrants*, we see a divergence between the sources in the trend towards more recent times, with the frequency of occurrence in the *Times* newspaper increasing while the frequency in parliamentary discourse decreases. In contrast, *refugees*, shows a closer pattern between the two sources although once again it is more frequent in the *Times* than Hansard up to the most

recent subcorpus when we actually see a small inversion. We may hypothesise that the incipient decrease in the *Times* reflects changing attitudes towards forced migration, but this requires a closer examination of the data.

2.4.3 Identifying c-collocates

In order to identify the c-collocates, in the first stage collocates were calculated for *refugees* in each decade in each of the two sub-corpora. Regarding the measure of collocation, it was essential to keep this the same for the two corpora and so the measure used had to be mutual information because this is what is available within the BYU interface. The span was set at 5L/R, the minimum frequency for collocates was set at 5 for both corpora and the minimum MI score was above 0. It should be noted that these are both relatively arbitrary measures, but for the purposes of comparison the key factor was that they were kept constant.

In the second stage, Excel was used to modify the lists as there is currently no dedicated tool for identifying c-collocates across corpora and the lists were too long to make manual matching time-efficient. I chose Excel for this case-study as it is relatively widely available and so I hope detailing the process may help others. Consistent collocates may be identified across two long lists by using the conditional formatting function to highlight duplicate texts between columns (assuming that columns correspond to collocates found in different sub-corpora). When looking at corpora where the size of the sub-corpora changes substantially over time, the shared collocates should be reported as a proportion of the collocates in the paired lists because we can expect that larger sub-corpora will yield greater numbers of shared collocates simply because there are more available, thus the comparison across time loses meaning if they are reported as raw figures. In order to track c-collocates across multiple lists, the COUNTIF function may be used to identify in how many columns (which in this case-study corresponds to collocates for decades) each term occurs.

There are two interrelated questions that we might pose regarding the consistency reference to *refugees* across time and discourse type:

1. To what extent are discourses consistent within one discourse type over a historical period? If present, what are these shared discourses?
2. To what extent are discourses shared between press and parliament? If present, what are these shared discourses?

These will be tackled briefly in the following two sections.

2.4.4 Investigating c-collocates

The first question we might ask is to what extent are collocates shared over time? The percentage of collocates of *refugees* which were shared between pairs of decades were calculated and are shown in Figure 2.4. What the sharing of collocates can tell us is whether the discourse/s surrounding *refugees* remain relatively stable and/or develop gradually over time. If there is a sudden drop in the number of shared collocates, then we would expect that to correspond to a marked shift in the discourse (although as always, this would only be an indicator and we would then need to delve into the corpus to analyse the texts).

[FIGURE 2.4 NEAR HERE]

The frequencies in Figure 2.4 suggest that the collocates were relatively stable over time. For both the *Times* and Hansard we see an increase in the proportion of shared collocates (note the trendlines (labelled as 'Linear' in the legend) go up), though this was more marked for Hansard. Shifting to difference, we might note that the *Times* consistently shows a larger number of shared collocates between years. This may be attributed to the fact that Hansard is more subject to variation as different political parties gain larger number of seats and therefore have more representation in the discourse overall.

To turn towards the second question posed above, the shared collocates between the *Times* and Hansard for decade were then identified and are reported in Figure 2.5 (the same vertical axis is maintained to ease comparison).

[FIGURE 2.5 NEAR HERE]

In Figure 2.5 we see a more marked shift over time. The trendline here is somewhat misleading because there is not a gradual increase, but it appears that the proportion of shared collocates increases substantially between 1900 and 1940. In the period from 1940 to 1999, the proportion then remains stable, suggesting that there is a consistent shared discourse between press and parliament.

However, it should still be noted that, overall, the greatest cohesion lies between the paired *Times* decades, followed by the Hansard decades, followed by each Hansard and *Times* pair for the same decade. Thus, overall, we can report that there is greater similarity within discourse types than across them. However, there is a flattening of difference between discourse types in the more recent time period and this would be an interesting focus for further investigation.

2.4.5 Patterns in the shared discourses

In the next stage, we move from the patterns of sharing, to illustrating how the shared items may be used as the basis for further similarity work. McEnery & Baker's (2017) study of prostitution in the seventeenth century pre-empted many of the issues involved in working with historical data found in this study. Like Gabrielatos & Baker (2008) they operationalised consistent collocates as those which occurred in seven of the ten decades under analysis. However, for the purposes of this study, the lower numbers of collocates for *refugees* in the early nineteenth century data and the missing decades in the *Times* data meant that this has to be modified further. Thus the items which were identified for further analysis in this case-

study were those which occurred in at least 50% of the decades. It should be clear that this proportion may be too low to talk about consistency and so the actual distribution is discussed further below.

Table 2.1 shows the collocates of *refugees* which occurred in at least half the decades analysis for Hansard. They have been grouped according to semantic themes which is a researcher-driven interpretation of what they are doing in the discourse (based on checking concordance lines, not abstracted dictionary-style meanings). For reasons of space, grammatical collocates (prepositions, determiners auxiliaries, conjunctions, modals) have not been included here.

[TABLE 2.1 NEAR HERE]

Table 2.1 shows that the semantic preference for quantification, as documented in contemporary migration discourse (e.g. Baker 2006) is strong over at least 5 decades of parliamentary discourse. We also see a pattern of deictic references to movement, and mentions of geographical locations at a national level. Alongside this, we have a concentration of items relating to assistance that may be offered to the people involved. In terms of problematizing and/or topicalizing *refugees*, we also see *problem* and *question* occurring repeatedly.

Table 2.2 shows the same data for the *Times*. As can be seen, there are a larger number of collocates here, partly because there was more similarity across the decades within the *Times* corpus (as shown above). The collocates were classified and the larger number led to a wider range of categories.

[TABLE 2.2 NEAR HERE]

As in Table 2.1, in this table we see the dominance of semantic fields relating to quantification, movement and nationality. We also see expansion of the people category to institutions who may be reacting to *refugees*. The items from the description category indicate the suffering of those described as *refugees*, pointing towards a continuous

sympathetic stance taken with this term. We might note that here the meta-references to the problematisation is not expanded as a category, and only *question* occurs. Similarly, the 'places' category remains relatively under-populated and does not indicate temporary locations (such as *camp* for instance).

Each of these groupings may constitute an avenue for further investigation in building up a picture of how *refugees* are positioned in discourse over time and across discourse types, as a way to understand the reflexivity between press and parliament.

2.5 Conclusions

In this chapter, I have tried to show both why we should consider similarity if we want to try and account fully for the discourses that we are analysing, and how we might approach it both with the tools available and various 'work arounds'. In the case-study examining collocates of the lexical item *refugees* over time, we have seen that looking at what is shared has confirmed the long-standing sympathetic stance of this lexical item. This continuous evaluation may account for the decline in use in the *Times*, given that the UK press has been increasing in its anti-immigration sentiment.

Through raising the topic of similarity and illustrating some of the methods we have available, I hope to place it into the standard set of practices with which we engage when do discourse analysis using corpora. In future, it may be that software packages will integrate features such as c-collocates and lockwords as standard tools which will help cement the relevance of similarity in research.

Although the case-study here has focused on similarity, the argument I wish to put forward is not just that we should have more research looking at what is shared but that the simple practice of 'looking both ways' will help us reflect on where our research direction is taking us, and to achieve a more 360 degree view of the discourse/s we are investigating.

Indeed, a singular focus on similarity may be counter-productive. As Bednarek & Caple (2017: 165) reflexively conclude after employing ProtAnt to investigate the discourses around cycling 'a focus on prototypicality and range may background variety to some extent', for instance, submerging differences between the newspapers in their corpora. As they go on, 'in the same way in which a focus on differences (which is generally more common in corpus linguistics) may create a "blind spot" (Taylor 2013: 83) a focus on similarities may do so, too' (Bednarek & Caple, 2017: 165-166).

Acknowledgements

I am very grateful to the University of Lancaster-based ESRC Centre for Corpus Approaches to Social Science (CASS) for allowing me access to the magnificent *Times* corpora while I was a visiting researcher there in 2016-2017.

References

Anthony, L. and Baker, P. 2015. ProtAnt: A tool for analysing the prototypicality of texts. *International Journal of Corpus Linguistics*, 20(3): 273-292.

Anthony, L. and Baker, P. 2017. ProtAnt (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Bachmann, I. 2011. "'Civil partnership – 'Gay marriage in all but name': a corpus-driven analysis of discourses of same-sex relationships in the UK Parliament', *Corpora* 6 (1): 77–105.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. Continuum.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P. 2011. ‘Times may change but we’ll always have money: a corpus driven examination of vocabulary change in four diachronic corpora’, *Journal of English Linguistics* 39, 65–88.

Bednarek, M., and Caple, H. (2017). *The Discourse of News Values: How News Organizations Create Newsworthiness'*. Oxford: Oxford University Press.

Davies, M. 2004. *BYU-BNC: The British National Corpus*. Available online, at: http://corpus.byu.edu/bnc/

Fitzsimmons-Doolan, S. 2009. ‘Is public discourse about language policy really public discourse about immigration? A corpus-based study’, *Language Policy* 8, 377–402.

Gabrielatos, C. and P. Baker. 2008. ‘Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005’, *Journal of English Linguistics* 36 (1): 5–38.

Germond, B., McEnery, T., and Marchi, A. (2016). The EU’s comprehensive approach as the dominant discourse: a corpus-linguistics analysis of the EU’s counter-piracy narrative. *European Foreign Affairs Review*, 21(1): 137-156.

Gries, S. Th. and Hilpert, M. 2008. ‘The identification of stages in diachronic data: Variability-based neighbour clustering’. *Corpora* 3(1): 59–81.

Hardie, A. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3): 380-409.

Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1): 97-133.

Malone, B. 20 August 2015. ‘Why Al Jazeera will not say Mediterranean ‘migrants’. *Al Jazeera*. http://www.aljazeera.com/blogs/editors-blog/2015/08/al-jazeera-mediterranean-migrants-150820082226309.html

Marchi, A. and C. Taylor. 2009a. ‘If on a winter’s night two researchers…a challenge to assumptions of soundness of interpretation’, *CADAAD Journal* 3 (1): 1–20.

McEnery, T. 2009. ‘Keywords and moral panics: Mary Whitehouse and media censorship’ in D. Archer (ed.) *What’s in a Word: Investigating Word Frequency and Keyword Extraction*, pp. 93–124. Farnham: Ashgate.

McEnery, A., and Baker, H. 2017. *Corpus Linguistics and 17th-century Prostitution: Computational Linguistics and History*. Bloomsbury.

McEnery, T. and A. Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meyer, C. 2002. English Corpus Linguistics: An Introduction. Cambridge and New York: Cambridge University Press.

Partington, A. 2009. ‘Evaluating evaluation and some concluding thoughts on CADS’ in J. Morley and P. Bayley (eds) *Corpus-assisted Discourse Studies on the Iraq Conflict: Wording the War*, pp. 261–304. London: Routledge.

Pearce, M. 2008. ‘Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine’, *Corpora* 3 (1): 1–29.

Rychly, Pavel and Adam Kilgarriff. 2007. “An efficient algorithm for building a distributional thesaurus”. *Proc ACL*. Prague Czech Republic. Available from http://www.kilgarriff.co.uk/Publications/2007-RychlyKilg-ACL-thesauruses.pdf

Scott, M. 1997. ‘PC analysis of key words – and key key words’, *System* 25(2): 233–45.

Scott, M. 2000. ‘Focusing on the text and its key words’ in L. Burnard and T. McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective*, volume 2, pp. 103–22. Frankfurt: Peter Lang.

Scott, M. 2001. ‘Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs’ in M.

Ghadessy, A. Henry and R.J. Roseberry (eds*) Small Corpus Studies and ELT: Theory and Practice*. Amsterdam and Philadelphia: John Benjamins.

Scott, M., 2016, *WordSmith Tools version 7*, Stroud: Lexical Analysis Software.

Scott, M. and C. Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam and Philadelphia: John Benjamins.

Seale, C, S. Boden, S. Williams, P. Lowe and D. Steinberg. 2007. 'Media constructions of sleep and sleep disorders: a study of UK national newspapers', *Social Science and Medicine* 65, 418–30.

Stubbs, M. 2005. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature* 14(1): 5-24.

Taylor, C. 2010. 'Science in the news: a diachronic perspective', *Corpora* 5 (2): 221–50.

Taylor, C. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8(1): 81-113.

Taylor, C. 2014. Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis. *International Journal of Corpus Linguistics*, 19(3): 368-400.